Clustering citation distributions for semantic categorization and citation prediction

Conference or Workshop Item

Clustering Citation Distributions for Semantic Categorization and Citation Prediction

Francesco Osborne[1], Silvio Peroni[2,3], Enrico Motta[1]

[1] Knowledge Media Institute, The Open University, Milton Keynes, UK
{francesco.osborne,e.motta}@open.ac.uk
[2] Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
silvio.peroni@unibo.it
[3] STLab, Institute of Cognitive Sciences and Technologies, CNR, Rome, Italy

Abstract. In this paper we present i) an approach for clustering authors according to their citation distributions and ii) an ontology, the *Bibliometric Data Ontology*, for supporting the formal representation of such clusters. This method allows the formulation of queries which take in consideration the citation behaviour of an author and predicts with a good level of accuracy future citation behaviours. We evaluate our approach with respect to alternative solutions and discuss the predicting abilities of the identified clusters.

Keywords: Semantic Web, Research Data, Bibliometric Data, Expert Search, Hierarchical Clustering, Data Mining, OWL, RDF, SPARQL, BiDO

1 Introduction

Exploring and analysing scholarly data [1] help to understand the research dynamics, forecast trends and derive new knowledge, which can be effectively represented by semantic technologies. Within this context, two important tasks are:

1) classifying authors according to a variety of semantic categories in order to facilitate querying, sharing and reusing such data in different context;
2) forecasting their career trends, allowing us to estimate their future citation behaviour.

In this paper we will present an innovative approach to address both tasks by exploiting author citation distributions.

Most of today systems for the exploration of academic data offer citations or citations-based indexes (e.g., h-index, g-index) as ranking metrics and provide interesting visualizations of citation distributions. However, they do not exploit many interesting features which can be derived by the analysis of citation distributions, such as: 1) the trend of the distribution within a certain time interval (e.g., it is steadily rising), 2) the timing of possible acceleration/deceleration (e.g., it started to rise much faster in the last 3 years), 3) the slope of the citation curve (e.g., every year it gains 20% more citations than the year before), 4) the shape of the citation curve (e.g., it is growing according to a logarithmic function), and 5) the estimated citation behaviour in the following years (e.g., authors with a similar pattern usually receive 200±50 citation in their 8th career years).

These features can support formulating queries that take in consideration the diachronic citation behaviour of authors. Examples are: “find all PhD students working in Semantic Web who exhibit a possible rising star pattern”, “find all the senior researchers who in their young years exhibited the same citation pattern as author X” or “find all the postdoc working in UK whose citations exhibit a positive trend in the last two years and are rising exponentially”.

Analysing the citation distributions can also foster a better understanding of the dynamics of an author career, since it makes possible to categorize different kinds of patterns and to study how they evolve. Moreover, it can allow us to forecast the future citation behaviour of research communities or organizations by studying the patterns of their members.

In this paper we present an approach for clustering authors according to their citation distributions, with the aim of extracting useful semantic information and producing statistical evidence about the potential citation behaviour of specific categories of researchers. In addition, we introduce an ontology, i.e., the *Bibliometric Data Ontology* (*BiDO*), which allows an accurate representation of such clusters (and their intended semantics) according to specific categories.

The rest of the paper is organized as follows. In Section 2, we discuss existing approaches for clustering authors and predicting future citations. Section 3 describes our approach for clustering authors’ citation distributions, while Section 4 illustrates BiDO and introduces the steps for associating the identified clusters to ontological categories. In Section 5, we evaluate our approach versus alternative solutions and discuss the predictive abilities of the identified clusters. Finally, in Section 6, we summarize the key contributions of this paper and outline future directions of research.

2 Related Work

Classifying entities associated to a time series is a common task that is traditionally addressed with a variety of clustering techniques [2]. Citation distributions and their mathematical properties have been carefully analysed in a number of empirical studies (e.g., [3]). However, while academic authors are often classified by community detection and clustering algorithms with the aim of identifying different kinds of research communities [4,5], no current model exploits clusters of citation distributions to classify researchers according to the features described earlier and estimate their future citation behaviour.

In the past, several works have been published about the identification of the factors that allow the prediction of future citations. Their analyses, and the related statistical models and machine learning techniques proposed for such predictions, are usually performed according to specific hypotheses: taking into consideration only articles of high-rated journals of a certain discipline; analysing only particular kinds of articles (e.g., clinical articles); choosing only multidisciplinary journals so as to increase the coverage (and the variability) of the research communities involved; and

so forth[2]. As a result, different starting hypothesis gave rise to different (even contrasting) discriminating factors and prediction models.

However, most of these works agree on the existence of two different and complementary kinds of factors:

- *intrinsic* factors, i.e., those related with the qualitative evaluation of the content of articles (quality of the arguments, identification of citation functions, etc.);
- *extrinsic* factors, i.e., those referring to quantitative characteristics of articles such as their metadata (number of authors, number of references, etc.) and other contextual characteristics (the impact of publishing venue, the number of citation received during time, etc.).

The use of intrinsic factors data can be very effective but also time consuming. They can be gathered manually by humans, e.g., through questionnaires to assess the intellectual perceptions of an article (as in peer review processes). For instance, in [7] the authors show how the editor's and reviewer's ratings (in the context of the *Journal of Cardiovascular Research*, http://cardiovascres.oxfordjournals.org) are good predictors of future citations.

The data of some intrinsic factors, such as the identification of citation functions (i.e., author's reasons for citing a certain paper), can also be gathered automatically with the aim of being used to provide alternative metrics for assessing or predicting the importance of articles through machine learning techniques (cf. [8]), probabilistic models (cf. [9]), and other architectures based on deep machine reading (cf. [10]),

However, these approaches use extrinsic factors, rather than intrinsic ones, for the analysis of the importance of articles, because of the time-consuming nature of the latter ones and the quick availability (usually at publication time) of most of the extrinsic-based data. In [11], Didegah and Thelwall investigate the extrinsic factors that better correlate with citation counts, identifying three factors as the best ones for such prediction: the impact factor of the journals where articles have been published, the number of references in articles, and the impact of the papers that have been cited by the articles in consideration. Other extrinsic factors identified in other studies are article length (in terms of printed pages) [12], number of co-authors [13], rank of author's affiliation [13], number of bibliographic databases in which a journal was indexed [14], proportion of the journal articles published that had been judged of high quality by some authoritative source [14], and price index [6]. Slightly different kinds of extrinsic factors were considered in Thelwall *et al.*'s work on altmetrics [15]. The authors analysed eleven different altmetrics sources and found that six of them were good predictors of future citations (i.e., tweets, Facebook posts, Nature research highlights, blog mentions, mainstream media mentions and forum posts).

3 Clustering Citation Distributions

In this section, we will present our approach for detecting clusters of researchers who share a similar citation distribution. We want to identify clusters characterized by citation distributions which represent the typical patterns of some categories of

[2] A good literature review of a large number of such approaches is available in [6].

authors, so that each cluster will suggest a common future behaviour. More formally, we want to subdivide the authors in sets, in such a way that the population of each set will remain homogenous with respect to the number of citations collected in the following years, i.e., the members of each cluster will have a similar number of citations also in the future.

Our approach takes as input the citation distributions of authors in a certain time interval and returns 1) a set of clusters with centroids that describe the most typical citation patterns, 2) a matrix associating each author with a number of clusters via a membership function, and 3) a number of statistics associated to each cluster for estimating the evolution of the authors in that cluster.

We cluster the citation distributions by exploiting a bottom-up hierarchical clustering algorithm. The algorithm takes as input a matrix containing the distance between each couple of entities and initially considers every entity as a cluster. It then computes the distance between each of the clusters, joining the two most similar clusters at each iteration. We adopt a single-linkage strategy by estimating the distance between two clusters C_1 and C_2 as the shortest distance between a member of C_1 and a member of C_2. The algorithm stops when it reaches a certain distance threshold *t*.

To obtain cluster sets that are fit for our purpose we must thus define accordingly 1) the metric to compute the distance between each couple of citation distributions and 2) a method to decide the threshold *t*.

It is possible to measure the distance between two time series by means of metrics such as the Euclidean distance or cosine similarity. Unfortunately both of these solutions have some shortcomings in this case. In fact, when using the Euclidean distance, covariates with the highest variance will drive the clustering process: a threshold value that allows clustering distributions of a certain scale (e.g., 200 citations) will also merge together perfectly valid clusters of minor scale (e.g., 20 citations). The distance based on the cosine similarity (e.g., the inverse minus one) will solve this problem since it is scale-invariant; unfortunately it would also cluster together distributions of completely different scale but with the same shape (e.g., [1,1,2] and [100,100,200]). Let us assume a couple of citation distributions *A* and *B* having both a total of *n* citations, and a different couple of them *C* and *D* with *m* citation each, *C* having the same distribution as *A*, and *D* the same as *B*. We want a distance that will yield $dis\{A,B\} = dis\{C,D\}$ (avoiding the covariate with the highest variance to drive the clustering) and also $dis\{A,C\} > 0$ (making scale a feature), and furthermore can be calculated incrementally (thus sparing processing time by stopping the computation over a threshold). A simple way to satisfy these three requirements makes use of a Euclidean distance normalized with the number of total citations of both distributions (similarly to [16]):

$$EU_n = \left(\frac{\sum_{i=0}^{n} (x_i - y_i)^2}{\sum_{i=0}^{n} (x_i)} + \frac{\sum_{i=0}^{n} (x_i - y_i)^2}{\sum_{i=0}^{n} (y_i)} \right) / 2 \qquad (1)$$

where x_i and y_i are the number of citations of the two distributions in the *i*-th year.

We also want to choose a threshold value *t* that will maximize the homogeneity of the cluster populations in the following years. We compute the homogeneity of a population with respect to citations using the Median Absolute Deviation (MAD). MAD is a robust measure of statistical dispersion [17] and it is used to compute the

variability of an univariate sample of quantitative data. It was first used by Gauss for determining the accuracy of numerical observations and it is defined as the median of the absolute deviations from the original data's median:

$$MAD = median_i\left(\left|x_i - median_j\left(x_j\right)\right|\right) \quad (2)$$

The procedure for computing the MAD consists in calculating the median of the n original data ($x_1,x_2,\ldots,x_j,\ldots,x_n$), computing the differences between each one of the n original values x_i and the median of the whole data distribution and finally computing the median of the previous differences. We preferred MAD to different solutions, such as standard deviation, for its robustness. In fact, standard deviation is too much influenced by outliers such as a few authors with a very high number of citations. Hence, we estimate the quality of a set of clusters in a certain year by computing the weighted average of their MAD:

$$MAD_{av} = \frac{\sum_{i=0}^{n}(MAD(c_i)\cdot dim(c_i))}{dim(c_i)} \quad (3)$$

where $MAD(c_i)$ and $dim(c_i)$ are respectively the MAD and the number of authors associated with the i-th cluster.

We set the threshold t by running the hierarchical algorithm with different t values and then selecting the threshold which yields clusters with the lowest average MAD_{av} in the following n years (n=10 in the herein presented evaluation). For characterizing completely the author space we compute the clusters for different intervals of time, e.g., 1-5, 1-10 and 1-15 career years, using a significant author sample (e.g., 5000). We then compute the memberships of all authors in our dataset with the centroids of the resulting clusters, so as to determine exactly how much a specific author is similar to each cluster centroid. For associating authors to clusters, we adopt the well known membership formula of the Fuzzy C-Mean algorithm [18], that is:

$$mem_k(x) = \frac{1}{\sum_{i=0}^{n}\left(\frac{dis(center_k,x)}{dis(center_j,x)}\right)^{2/(m-1)}} \quad (4)$$

where $mem_k(x)$ is the membership value of author x with cluster k, $dis(center_i, x)$ the distance between x and the centroid of cluster i, and m is a constant for modulating the level of cluster fuzziness (m=2 in the prototype).

Finally we analyse the distribution of each cluster population with respect to the number of citations received in the following years, in order to extract statistical evidence about their future behaviour. As mentioned before, standard deviation is severely influenced even by few outliers, making it hard to use the mean on the full population as a predictor. Hence, for each year we automatically select a percentage p of the population (e.g., 90%) in the most populated area of the distribution and compute its interval of citations (e.g., 40-80), mean (e.g., 45) and standard deviation (e.g., 14). Technically, we do so by computing the number of authors who fall into different ranges of citations, ordering those categories in decreasing order and then selecting the authors from subsequently smaller categories until the percentage of authors selected is equal to p. The citation interval, mean and standard deviation of this sample produce accurate, intuitive and statistically sound predictions which are more resilient to outliers.

Intuitively, some categories of authors are too mundane to suggest a common future behaviour, and may be used only for classification purposes. Hence, in this phase we care especially about the "uncommon signature" that points to particularly homogenous population of authors. Figure 1 shows the distributions of authors in their seventh career year associated to some clusters detected by analysing their first five career years (the dashed line refers to the overall distribution). Clusters *C29* and *C30* are associated with a very specific citation patterns and thus their distributions have a small kurtosis and point to two narrow categories of authors who normally receive a relatively low number of citations. Clusters *C25* and *C28* are also quite homogeneous and represent two distinct populations of more frequently cited authors. Naturally, the homogeneity of the population associated with a cluster will decrease in the following years and so will the accuracy of the predictions.

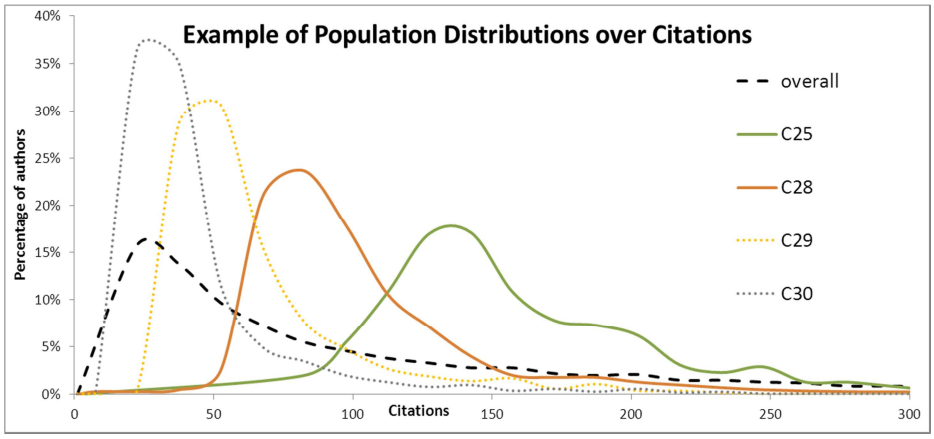

Figure 1. Percentage of authors vs. number of their citations in their 7th career year. The clusters were derived by the citations received over the first 5 career years.

4 An ontology for describing bibliometric data

Having a model developed according to a well-known format (such as OWL) for enabling the classification of authors and journals according to bibliometric data is crucial to allow one to query, share and reuse such data in different context, e.g., for providing smart visualisation of bibliometric data for sense-making activities and for enabling automatic reasoning on them.

However, bibliometric data are not simple objects, since they are subject to the simultaneous application of different variables. In particular, one should take into account at least:

- the *temporal association* of such data to entities, in order to say that a particular value, e.g., the fact that an article has been cited 42 times, was associated to such article only for a time period;
- the particular *agent who provided* such data (e.g., Google Scholar, Scopus, our algorithm), in order to keep track of the way data evolve in time according to particular sources;

- the *characterisation* of such data in at least two different kinds, i.e., numeric bibliometric data (e.g., the standard bibliometric measures such as h-index, journal impact factor, citation count) and categorial bibliometric data (so as to enable the description of entities, e.g., authors, according to specific descriptive categories).

The *time-indexed value in time* (*TVC*) ontology design pattern [19] seems to be a good starting model for the development of an ontology for bibliometric data, since TVC's entities enable the precise description of all the aforementioned variables: time, responsible agent and kinds of data.

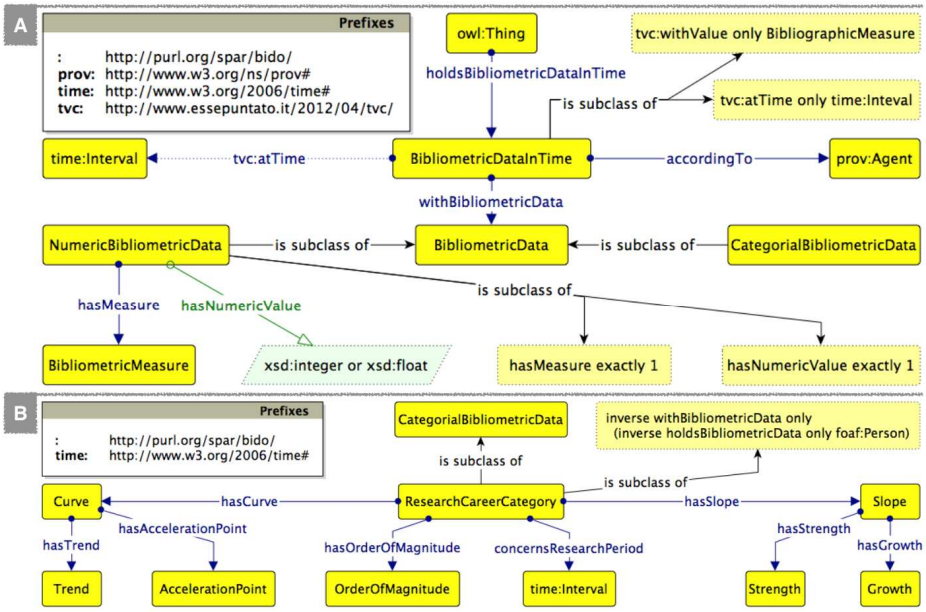

Figure 2. *A:* the core module of BiDO, describing generic bibliometric data with their characterising variables. *B:* the module modelling a particular kind of categorial bibliometric data, i.e., the research career categories, according to the main dimensions used by the algorithm in Section 3.

Starting from TVC, we have created the *Bibliometric Data Ontology* (*BiDO*, available at http://purl.org/spar/bido), i.e., a modular OWL 2 ontology that allows the description of bibliometric data of people, articles, journals, and other entities described by the SPAR Ontologies (http://purl.org/spar) in RDF.

The core module of the ontology, shown in Fig. 2.A, allows us to describe any entity and the related bibliometric data (through the property *holdsBibliometricDataInTime*) at a certain time (i.e., *tvc:atTime*, a property defined by the imported TVC ontology for specifying temporal instants or intervals) and according to a certain agent (through the property *accordingTo*, which is a sub-property of *prov:wasAttributedTo* and allows us to indicate the agent responsible for such bibliometric data). In addition, BiDO imports PROV-O [20] for adding provenance data about the activities related to the creation of such bibliometric data.

Two alternative kinds of bibliometric data are specifiable (through the property *withBibliometricData*) in BiDO: numeric and categorial bibliometric data. Numeric bibliometric data are those characterised by a certain integer or float value related to a particular bibliometric measure. Some of these measures – i.e., *h-index*, *author citation count*, *e-index*, and *journal impact factor* – are available in a particular module of BiDO responsible for describing the most common bibliometric measures.

We have developed an additional module of BiDO that extends the class *CategorialBibliometricData* of the core module with specific categories describing the research career of people, in order to address the mapping of the clusters identified by the algorithm presented in Section 3 with specific facets. As shown in Fig. 2.B, these facets are described by the class *ResearchCareerCategory*, which is characterised by four specific dimensions that have been used by our algorithm to cluster citation data:

- the *research period* considered, i.e., the interval of research years that the algorithm is taking into consideration (e.g., the first 5/10 years);
- the *curve*, i.e., the specific shape proper to the clusters identified by the algorithm, which is characterised by a trend (flat/increasing/decreasing) and, in the latter two cases, by an acceleration or deceleration point (none or premature, median, overdue acceleration/deceleration);
- the *slope* of such curve, in terms of strength (low/moderate/high) and kind of growth (linear/polynomial/exponential/logarithmic);
- the *order of magnitude*, which categorises the number of citations received in the considered period according to a uniform model of common-sense estimation [21], which describes intervals of half-order of magnitude – i.e., "[0,1)", "[1,3)", "[3,9)", "[9,27)", "[27,81)", "[81,243)", "[243,729)", etc.

The combinations of all these values related to the aforementioned dimensions have been used to define all the possible descriptive categories of research career of people as instances of the class *ResearchCareerCategory*.

Even if we did not define a particular category for each cluster found by the algorithm – rather, more clusters can be described by the same category –, we have defined an algorithmic procedure to determine the association between the cluster centroids and the categories described by the ontological model. For instance, let us consider the centroid "[31.3, 46.1, 52.8, 55.3, 60.8]"[3] of one of the clusters detected by our algorithm according to the first 5 years of research career. The related dimensions are identified in the following way:

- *order of magnitude*: we sum the values of the cluster centroid and select the interval containing such sum, i.e., "[243,729)";
- *curve trend*: the linear regression of the centroid is calculated, and then its slope is divided by the mean of all the centroid values. If the result of such division is greater than 0.05, then we have an increasing trend (which is the case of our example, since that value is 0.14), if it is less than -0.05 we have a decreasing trend, otherwise we have what we can approximately consider a flat trend;

[3] The five values of the centroid identify the number of citations that have been received during the five years of the research period considered.

- *curve acceleration*: the ratio of the slopes of the linear regressions of series *k-n* and *1-k* (for each *k* between 2 and *n* - 1, where *n* is length of the list of values defining a cluster centroid) is calculated, in order to identify in which year (i.e., *k*) the acceleration or deceleration (this is the case of our example) happens, if any. Then, the acceleration/deceleration is considered premature if $k \leq \lceil n/3 \rceil$ (as in our example), overdue if $k \geq \lceil 2n/3 \rceil$, and median otherwise;
- *slope strength*: the linear regression of the centroid is calculated, its slope is divided by the mean of all the centroid values, and then we calculate the absolute value *s* of this division. We say that the slope strength is low if *s<0.25* (as in our example), high if *s>0.45*, and moderate otherwise;
- *slope growth*: by means of the least squares method, we create the four functions (one linear, one polynomial, one exponential and one logarithmic) that best match with the cluster centroid. Then we compare the centroid data with such functions through Wilcoxon's non-parametric test for matched data and choose the best fitting function (logarithmic in our example).

Following these steps, the example cluster we considered is mapped in the following category:

```
:increasing-with-premature-deceleration-and-low-logarithmic-slope-in-[243,729)-5-
years-beginning a :ResearchCareerCategory ;
  :hasCurve [ a :Curve ;
    :hasTrend :increasing ; :hasAccelerationPoint :premature-deceleration ] ;
  :hasSlope [ a :Slope ; :hasStrength :low ; :hasGrowth :logarithmic ] ;
  :hasOrderOfMagnitude :[243,729) ;
  :concernsResearchPeriod :5-years-beginning .
```

Thus, combining the results of our clustering algorithm with BiDO it is possible to associate authors with specific categories describing their research career as follows:

```
ex:john-doe :holdsBibliometricDataInTime [
  a :BibliometricDataInTime ;
  tvc:atTime [ a time:Interval ; time:hasBeginning :2014-07-11 ] ;
  :accordingTo [ a fabio:Algorithm ;
    frbr:realization [ a fabio:ComputerProgram ] ] ;
  :withBibliometricData
    :increasing-with-premature-deceleration-and-low-logarithmic-slope-in-
[243,729)-5-years-beginning .
```

The RDF descriptions of such bibliometric data make easier to query them with standard languages such as SPARQL, in order to retrieve, for instance, all the authors that in the first 5 years of their research career had a citation behaviour pattern like that described by the aforementioned category.

5 Evaluation

We evaluated our method on a dataset of 20000 researchers working in the field of computer science in the 1990-2010 interval. This dataset was derived from the database of Rexplore [1], a system that combines statistical analysis, semantic technologies and visual analytics to provide support for exploring scholarly data, and integrates several data sources (Microsoft Academic Search, DBLP++ and DBpedia).

In particular we wanted to show that the normalized Euclidean distance introduced in Section 3 works better than other choices for the task of clustering citation distributions. Hence, we compared three metrics: the normalized Euclidean distance (label NEU), the Euclidean distance (EU) and the distance based on the cosine similarity (CO). We measured the quality of the produced set of clusters in a certain year by their MAD_{av}, as in Formula (3).

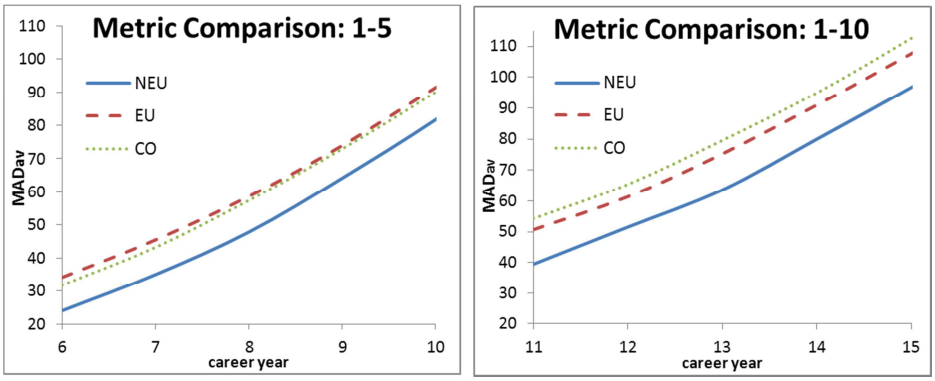

Figure 3. Comparison between NEU, EU and CO applied on the first five and ten career years according to their MAD_{av} in the following five years.

Figure 3 shows the performance of the three techniques when clustering the first five and ten career years. In all cases the normalized version of the Euclidean distance performs much better than the other solutions, being characterized by a smaller MAD_{av} value, e.g., a smaller degree of dispersion. CO performs slightly better than EU in the 1-5 years interval while EU performs better than CO in the 1-10 years interval. Analogous results were obtained by considering the weighted average of standard deviation rather than MAD_{av}.

Career year	C18 (1.4%)		C22 (2.5%)		C25 (2.7%)		C28 (2.3%)		C29 (8.8%)	
	range	mean±s.d.	range	mean±s.d.	range	mean±s.d.	range	mean±s.d.	range	mean±s.d.
6	420-800	567±98	160-280	209±34	100-180	129±25	60-100	72±14	40-60	39±9
7	440-960	610±120	160-320	225±45	100-200	138±30	60-120	79±18	40-80	45±14
8	440-1020	650±137	160-400	246±58	100-260	158±45	60-160	90±26	40-100	50±18
9	440-1260	699±186	160-440	269±74	100-340	187±68	60-200	104±37	40-120	57±25
10	480-2940	751±411	160-500	292±85	100-400	211±82	60-280	125±57	40-160	68±35
11	480-2480	826±336	180-660	331±112	100-520	241±100	60-540	155±103	40-200	82±47
12	480-3520	914±467	180-860	370±151	100-640	270±126	60-440	166±96	40-260	97±60

Table 1. Range of citations and mean citations in subsequent career years predicted with 75% accuracy for authors associated with clusters detected in the 1-5 career year interval. In parenthesis the percentage of authors in each cluster.

Our approach yields a number of clusters with different prediction capabilities. We can suggest a narrower or larger interval of predicted citations for increasing or lowering the precision of our predictions. Table 1 shows some example of predictions that yield 75% accuracy. For example we are able to suggest with 75% precision that 2.5% authors in Computer Science associated with cluster *C22* will have 225±45 average citations in their seventh career years (with a minimum number of citations equal to 160 and a maximum one equal to 320).

The left panel of Figure 4 shows the citation distributions of the centroids of the cluster in Table 1 and the algorithm predictions. Even if the predictions become less accurate in time, however they still can give a fair idea of the kind of potential citation behaviour of the authors. Moreover, these predictions are particular valuable for forecasting the future citation behaviour of an organization or research communities. In fact, while it is relatively hard to foresee a single author's citation behaviour (e.g., she/he may be an outlier), it is much easier to compute the predicted citations of a group of authors since in a large sample statistical fluctuations have a smaller weight.

Finally, the right panel of Figure 4 shows the evolution of some the main clusters in terms of average citations of the associated authors. We can notice that our approach allows a very good coverage of the possible career trajectories, from the most modest to the outstanding ones. This variety of patterns allow also for a very fine-grained semantic classification of researcher careers.

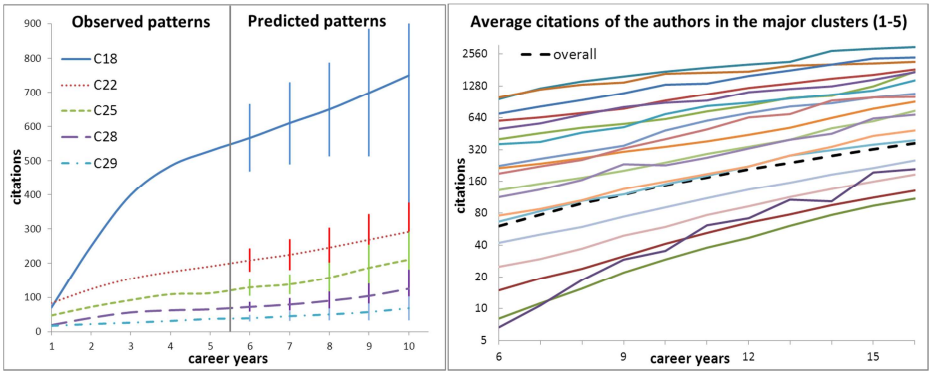

Figure 4. Left Panel: the citation distributions of the centroids of the clusters in Table 1 and the resulting predictions (the error bars represent the standard deviations of the predicted citations). Right Panel: the evolution in term of average number of citations of the authors associated to the main clusters in the 1-5 interval.

6 Conclusion

In this paper, we presented a novel approach for clustering author's citation distributions, with the aim of 1) classifying authors with a variety of semantic facets, and 2) forecasting the citation behaviour of categories of researchers. We also introduced the Bibliometric Data Ontology, a.k.a. BiDO, which is an OWL ontology that allows an accurate representation of such semantic facets describing people's research careers. In addition, we showed that our approach outperforms other solutions in terms of population homogeneity and is able to categorize a variety of career trajectories, some of which allow predicting future citations with fair accuracy.

For the future we plan to augment the clustering process with a variety of other features (e.g., research areas, co-authors), to extend BiDO in order to provide a semantically-aware description of such new features, and to make available a triplestore of bibliometric data linked to other datasets such as Semantic Web Dog Food and DBLP.

References

1. Osborne, F., Motta, E., Mulholland, P.: Exploring Scholarly Data with Rexplore. In Proceedings of the ISWC 2013: 460-477. (2013)
2. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. Journal of Intelligent Information Systems, 17(2-3), 107–145. (2001)
3. Redner, S.: How popular is your paper? An empirical study of the citation distribution. The Physics of Condensed Matter Journal, 4(2), 131–134. (1998)
4. Ding, Y.: Community detection: topological vs. topical. Journal of Infometrics, 5(4). (2011)
5. Osborne, F., Scavo, G., Motta, E.: Identifying diachronic topic-based research communities by clustering shared research trajectories. In The Semantic Web: Trends and Challenges (pp. 114-129). Springer International Publishing. (2014)
6. Onodera, N., & Yoshikane, F.: Factors affecting citation rates of research articles: Factors Affecting Citation Rates of Research Articles. Journal of the Association for Information Science and Technology. (2014)
7. Opthof, T.: The significance of the peer review process against the background of bias: priority ratings of reviewers and editors and the prediction of citation, the role of geographical bias. Cardiovascular Research, 56(3), 339–346. (2002)
8. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In Proceedings of the EMNLP 2006: 103–110. Stroudsburg, Pennsylvania, USA. (2006)
9. Dietz, L., Bickel, S., Scheffer, T.: Unsupervised prediction of citation influences. In Proceedings of ICML 2007: 233–240. (2007).
10. Di Iorio, A., Nuzzolese, A. G., Peroni, S.: Towards the automatic identification of the nature of citations. In Proceedings of SePublica 2013. (2013)
11. Didegah, F., Thelwall, M.: Determinants of research citation impact in nanoscience and nanotechnology. Journal of the American Society for Information Science and Technology, 64(5): 1055–1064. (2013)
12. Falagas, M. E., Zarkali, A., Karageorgopoulos, D. E., Bardakas, V., Mavros, M. N.: The Impact of Article Length on the Number of Future Citations: A Bibliometric Analysis of General Medicine Journals. PLoS ONE, 8(2), e49476. (2013)
13. Antonakis, J., Bastardoz, N., Liu, Y., Schriesheim, C. A.: What makes articles highly cited? The Leadership Quarterly, 25(1), 152–179. (2014)
14. Lokker, C., McKibbon, K. A., McKinlay, R. J., Wilczynski, N. L., Haynes, R. B.: Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. BMJ, 336(7645), 655–657. (2008)
15. Thelwall, M., Haustein, S., Larivière, V., Sugimoto, C. R.: Do Altmetrics Work? Twitter and Ten Other Social Web Services. PLoS ONE, 8(5), e64841. (2013)
16. Ning, T.: Computing correlation integral with the Euclidean distance normalized by the embedding dimension. In Proceedings of ICSP 2008: 2708–2712. (2008)
17. Hoaglin, D. C., Mosteller, F., Tukey, J. W.: Understanding robust and exploratory data analysis (Vol. 3). New York: Wiley. (1983)
18. Bezdek, J. C., Ehrlich, R., Full, W.: FCM: The fuzzy c-means clustering algorithm. Computers and Geosciences, 10(2), 191–203. (1984).
19. Peroni, S., Shotton, D., Vitali, F.: Scholarly publishing and linked data: describing roles, statuses, temporal and contextual extents. In Proceedings of i-Semantics 2012: 9–16. (2012).
20. Lebo, T., Sahoo, S., McGuinness, D.: PROV-O: The PROV Ontology. W3C Recommendation, 30 April 2013. World Wide Web Consortium. (2013)
21. Hobbs, J. R., Kreinovich, V.: Optimal choice of granularity in commonsense estimation: Why half-orders of magnitude? International Journal of Intelligent Systems, 21(8), 843–855. (2006)